

(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

EP 0 789 298 A1

(12)

## EUROPEAN PATENT APPLICATION

(43) Date of publication:

13.08.1997 Bulletin 1997/33

(51) Int. Cl.<sup>6</sup>: G06F 9/38

(21) Application number: 97101810.6

(22) Date of filing: 05.02.1997

(84) Designated Contracting States:  
DE GB

(30) Priority: 07.02.1996 US 597784

(71) Applicant: Hewlett-Packard Company  
Palo Alto, California 94304 (US)

(72) Inventors:

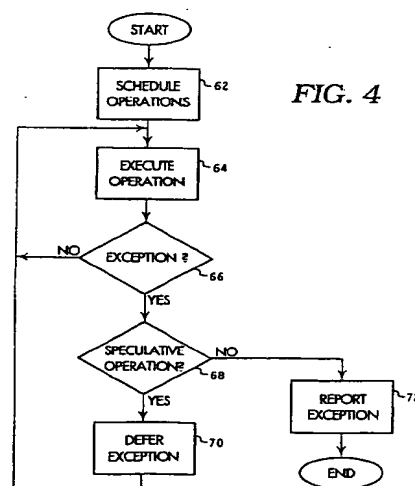
- Amerson, Frederic C.  
Santa Clara, CA 95051 (US)
- Rau, Ramakrishna B.  
Los Altos, CA 94024 (US)

- Gupta, Rajiv  
Los Altos, CA 94024 (US)
- Schiansker, Michael S.  
Los Altos, CA 94024 (US)
- Kathail, Vinod K.  
Cupertino, CA 95014 (US)
- Worley, William S., Jr.  
Denver, CO 80206 (US)

(74) Representative: Schoppe, Fritz, Dipl.-Ing.  
Patentanwalt,  
Postfach 71 08 67  
81458 München (DE)

## (54) Method and system for processing speculative operations

(57) A method for supporting speculative execution includes designating operations as speculative or non-speculative, and then deferring exceptions generated by speculative operations while immediately reporting exceptions by non-speculative operations. If a speculative operation uses a result of a speculative operation that has generated an exception, the exception is propagated. Deferred exceptions are detected and reported using a check operation either incorporated into a non-speculative operation or inserted as a separate check operation. A system for supporting speculative execution includes a functional unit for recognizing a speculative operation and deferring any exceptions generated by such an operation. The functional unit may defer an exception by storing information indicating an error has occurred in the register file. To check for deferred exceptions, the functional unit then reads the register file. If an exception is detected, then the exception is processed and one or more of the speculative operation are re-executed (in a non-speculative mode) where necessary to process the exception.



EP 0 789 298 A1

**Description****REFERENCE TO PARENT CASE**

5 This is a continuation-in-part of patent application Serial No. 08/324,940, entitled "METHOD AND SYSTEM FOR SUPPORTING SPECULATIVE EXECUTION" filed October 18, 1994, which is hereby incorporated by reference.

**FIELD OF THE INVENTION**

10 This invention generally relates to code optimization in computer scheduling and more particularly the architectural support for such optimization.

**BACKGROUND OF THE INVENTION**

15 The performance of a computer system may be enhanced by optimizing the code of a computer program so that the computer can execute the program more quickly. One of the steps in optimizing a program is a process called scheduling. Scheduling is a process where the series of computer operations that comprise a program are organized for execution. During the scheduling process, operations of the program may be arranged, eliminated, or moved to make the program run more efficiently for a particular CPU design. Generally, there are two forms of scheduling:  
 20 dynamic scheduling performed by the hardware during execution of a program, and static scheduling performed by a compiler before execution. Either of these techniques, or a combination of both, may be used to schedule operations in a computer program for processing by a computer system.

A computer program consists of a series of instructions to be carried out by a central processing unit (CPU) in the computer system. A typical program is written in a high level language and then compiled into a series of instructions  
 25 compatible with the instruction set architecture of the CPU. A program, however, may also be directly written in "machine language" according to the instruction set architecture of the computer. The instruction set architecture defines the format or encoding of operations, including operators and operands in an instruction. Depending on the structure of the CPU and the scheduling techniques involved, each instruction may have one or more operations. An operation includes an operator encoded in an opcode representing functions such as add, subtract, load, store, branch, etc. Additionally, an operation identifies the inputs and the results of the operation. To accomplish this, the operation typ-  
 30 ically includes a code identifying the location such as a register of an operand or operands. It is these operations that are organized for execution by the CPU using the optimization techniques.

There are different levels of optimization. One level of optimization is local optimization where code within a straight-line code fragment or "basic block" is manipulated to run more efficiently. Examples of local optimization tech-  
 35 niques are common subexpression elimination and constant propagation. Another level of optimization is global optimization which includes extending local optimization techniques across conditional branches in a program and further includes transformations for optimizing loops. One form of global optimization is code motion. An example of code motion is removing code from a loop that computes the same value each iteration of a loop. A third level of optimization is machine dependent optimization. Machine dependent optimization involves manipulation of code to take advantage  
 40 of specific architectural attributes of the CPU. For example, if the CPU has a pipelined functional unit for executing instructions concurrently, then code can be reordered to improve pipeline performance.

To optimize a program, a compiler may move code above a conditional branch in a scheduling process called spec-  
 45 ulative code motion. Speculative code motion refers to the movement of an instruction above a conditional branch that controls its execution. The execution of a "speculative" instruction may be referred to as a speculative or anticipatory execution because the instruction is executed before it is known whether the instruction will actually be used in the program. Speculative code motion can enhance instruction level parallelism. Because many instructions have a long latency, meaning they take several clock cycles to execute, it is advantageous to execute an instruction speculatively. The delay that an instruction would otherwise cause can be minimized by issuing the instruction in advance. Specula-  
 50 tive code motion may also be useful in other optimizations such as redundancy elimination.

While speculative execution has the potential for enhancing performance, it has not been effectively implemented due to the problems in dealing with errors in speculative instructions. If a speculative instruction generates an error, the error should be deferred and should only be reported if the speculative instruction is actually used in the program in normal sequence. Otherwise, errors that would not have been reported in the original program would be reported, and the behavior of the program would be changed.

55 Existing optimization techniques have not effectively addressed the problems with moving code across conditional branches. One way to address the error problem is to prevent speculative code motion for instructions that may generate an error. This solution has serious limitation because this significantly limits the number of operations which may be moved. Another potential solution is to ignore all errors. This solution is unsatisfactory because it prevents error handling or reporting.

Existing computer systems do not effectively take advantage of speculative execution, and specifically, do not support deferring and re-executing speculative operations to improve performance.

## SUMMARY OF THE INVENTION

To address the drawbacks and limitations of existing optimization techniques, the invention provides a method and system for supporting speculative execution. In one embodiment, the method includes providing two versions of the operations in the instruction set, a speculative version and a non-speculative version. Operations that are moved above conditional branches are encoded as speculative operations. Both speculative and non-speculative operations are then executed by a CPU. If a non-speculative operation generates an exception, the exception is handled or reported immediately.

In contrast, exceptions generated by speculative operations are deferred to avoid unnecessary exception processing. If a speculative operation generates an exception, then the exception is deferred. Deferring an exception can include propagating an exception when another speculative operations uses the result of a speculative operation that has generated an exception, or has propagated an exception. If the result of a speculative operation is used non-speculatively, then it is checked to determine whether an exception has been deferred.

In a second embodiment, the method includes providing natively speculative operations in the instruction set. These natively speculative operations are treated as speculative operations for scheduling and for deferring exceptions. However, these operations do not need to be identified as speculative operations by encoding them as such in the opcodes. Rather, the CPU treats a predetermined set of operations as natively speculative. To enhance the compatibility of the CPU with existing programs, the CPU can also include two modes of operation: a first mode where the predetermined set of operations are natively speculative, and a second mode where the predetermined set of operations are non-speculative. As in the first embodiment above, exceptions for speculative operations are deferred while exceptions for non-speculative operations are reported immediately.

A system for supporting speculative execution may be implemented in a CPU including a functional unit and register file. During execution of operations in a program, the functional unit of the CPU recognizes whether an operation is speculative or non-speculative. If a speculative operation generates an exception, the functional unit stores information in a register in the register file indicating that an exception has been deferred. In one implementation, the registers in the register file may include an error tag for identifying whether an exception has been deferred. A non-speculative operation may be used to check for deferred exceptions. In the process of executing such an operation, the functional unit checks the register file to determine whether an exception has been deferred. If an exception has been deferred, the functional unit reports the exception.

The method and System for providing support for speculative execution provide several advantages. Exceptions generated by speculative operations are ignored unless the speculative operation is actually used in the program. As a result, the computer system achieves the performance gains provided by speculative execution without losing full error detection and reporting capabilities. Employing the approach of one embodiment, speculative execution may be implemented in an instruction set architecture where speculative and non-speculative versions are provided for several operations. If, however, such an approach is not possible in a particular instruction set architecture, speculative execution can be supported by treating several operations as natively speculative without changing or specially encoding opcodes.

This alternative approach is especially useful in retrofitting an existing instruction set architecture to support speculative execution. To retrofit an existing system, the hardware of a CPU can be modified to support speculative execution while maintaining compatibility with programs written for an existing instruction set architecture. The instruction set is not materially changed, but instead, the hardware semantics of the operations are changed to add support for deferring and checking for deferred exceptions generated by speculative operations. In this approach, the CPU supports two modes of operation: a first mode where operations are non-speculative; and a second mode where some operations are treated as natively speculative. The second mode differs from the first in that exceptions are deferred for natively speculative operations. By supporting these two modes of operation, the CPU is compatible with the binary form of programs compiled for an existing instruction set architecture whether or not the programs have been optimized using speculative code motion.

Further advantages and features of the invention will become apparent to those skilled in the art from the following description and accompanying drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

- Fig. 1 is a block diagram of a computer system in which an embodiment of the invention may be implemented.
- Fig. 2 is a block diagram of a central processing unit in which an embodiment of the invention may be implemented.
- Fig. 3 is a diagram of an instruction having multiple operations.
- Fig. 4 is a flow diagram of a method for supporting speculative execution in an embodiment of the invention.

Fig. 5 is a diagram used to describe how fix-up code is generated to support re-execution of operations.

## DETAILED DESCRIPTION

5 An embodiment of the invention generally provides a method and system for supporting speculative execution. The following description begins with a general explanation of speculative execution. Next, systems for supporting speculative execution are described. Finally, a method for supporting speculative execution is described.

Speculative code motion can be used effectively in CPUs having concurrent processing capability. A CPU can execute a series of operations more quickly by issuing separate operations at the same time or overlapping the execution  
10 of operations. A typical CPU executes an operation in stages. In a CPU having more than one functional unit for executing operations, the CPU can issue multiple operations and perform the same stage for multiple operations simultaneously. In a pipelined processor, the CPU includes a pipeline for processing each stage in an operation like an assembly line. Pipelining reduces overall processing time of a program by enabling the processor to execute different stages of several instructions simultaneously. A CPU can include either multiple functional units or pipelining, or can  
15 combine both features.

In concurrent processing CPUs, speculative code motion can optimize performance by addressing the latency problem. Latency is the time required for the CPU to complete an operation. An operation with long latency can stall or delay execution of a program while the CPU waits to complete an operation before beginning another. Without speculative code motion the CPU must evaluate a branch condition before issuing any of the operations after the branch. By  
20 moving an operation above the conditional branch, the CPU can issue the operation in advance as a speculative operation. In a CPU with multiple functional units or pipelining, the CPU can process the speculative operation concurrently with other operations. If the branch where the operation originated is followed, then the operation will be either partially or entirely completed by the time the program reaches the operation's original position. Speculative code motion addresses the latency problem by reducing or even eliminating the stall in execution of an operation. Speculative code  
25 motion is particularly useful in processors with concurrent processing capability because the CPUs in such systems have the capacity to process multiple operations at once.

The process of executing speculative instructions is called speculative execution. Speculative code motion may be implemented in a variety of scheduling techniques to provide support for speculative execution. Speculative code motion may be used in connection with multiple functional unit or pipelined CPUs to exploit parallelism and thereby optimize a program. Speculative code motion can be used in conjunction with partial redundancy elimination to optimize a  
30 program for virtually any type of CPU. In addition to scheduling support, speculative execution also requires architectural support.

To implement speculative execution, two primary issues must be addressed: speculative code motion, and exception detection and reporting. In this context, speculative code motion refers to the scheduling technique used to move operations above their associated branches. Exception detection and reporting relates to scheduling in view of the possibility that a speculative operation may generate an error. Additionally, it refers to detecting and reporting errors generated by speculative operations. Because these issues are closely related to scheduling, they are discussed first in the context of scheduling to support speculative execution. Next, a preferred method of exception detection and reporting is discussed.  
35

Before a program is executed, a scheduler optimizes the program by using speculative code motion and perhaps other optimization techniques to decrease processing time of the program. The scheduler may be dynamic or static, or may include a combination of compiler and hardware scheduling. Any of a number of well known scheduling techniques may be used, including but not limited to trace scheduling, modulo scheduling, and enhanced pipelining. Regardless of the scheduling technique used, the scheduler should include support for speculative code motion.  
40

Speculative code motion includes moving an operation from its home basic block to a previous basic block. A basic block is a straight line sequence of operations followed by a single branch. The home basic block of an operation is the basic block in which it resides in the original program. The previous basic blocks for a given basic block are all the basic blocks that can branch to the given basic block or that sequentially precede the basic block. During scheduling, an operation may be moved across a branch to a set of previous basic blocks. The operation is moved such that the operation  
45 will always be executed before any use of its result will occur. To satisfy this requirement, the operation may be moved to multiple preceding basic blocks. This general technique of speculative code motion may be used in a variety of scheduling methods.

One method for scheduling across basic blocks is superblock scheduling. A superblock is a block of operations where control can only enter from the top but may exit at one or more points. Superblock scheduling consists of two  
50 steps: dependence graph construction and list scheduling. The dependence graph represents the control and data dependencies between operations in a superblock. Using control dependencies, the scheduling technique addresses restrictions on speculative code motion. In this context, the two primary restrictions on moving an operation above a branch are 1) the result of the operation is not used before it is redefined when the branch is taken; and 2) the operation will not cause an exception that alters the results of the program when the branch is taken.

Different scheduling techniques using the superblock model may address these general restrictions on speculative code motion. For most scheduling techniques, the first restriction can be overcome using compile-time renaming transformations. After control dependencies are eliminated, list scheduling using the dependence graph, instruction latencies, and resource constraints are performed to determine how operations are scheduled. The problem of exceptions may be handled differently depending on the scheduling technique.

An exception is an error such as arithmetic underflow or overflow, an undefined instruction, or a memory access error. When an operation causes an exception in a non-speculative system, the exception is typically reported immediately, and the CPU branches to an exception handling routine of the operating system. If speculative operations immediately reported an exception, then the CPU would have to perform potentially unnecessary and semantically incorrect error handling for speculative operations that may ultimately not be in the execution path of the program.

The preferred way of dealing with exceptions in speculative operations is to defer reporting of the exceptions. Using this technique, the second restriction on moving operations above conditional branches listed above is eliminated.

The exception is only reported if the operation is actually executed in its home basic block. By deferring such exceptions, unnecessary exception handling is avoided. The computer system provides enhanced performance and preserves the semantics of the original program.

If the result of a speculative operation that generates an exception is used as an input to a second speculative operation, then the exception is propagated to the second operation. An exception will be propagated through any speculative operations that use the result of a speculative operation generating an error. These propagated exceptions will only be reported if the exception is propagated to an operation that is actually used in the execution path of the program.

To support deferred reporting of exceptions, a system according to an embodiment of the invention includes a means for deferring the exception. An exception may be deferred using a buffer or a tag in the register to keep a record that an error has occurred. The system also includes a means for reporting a deferred exception. The home block of a speculative instruction includes an operation for checking whether an exception has occurred. This operation may be an additional check operation inserted in the home block or may be incorporated into a non-speculative operation in the home block. An implementation of these aspects of a computer system are described below.

As an overview, Fig. 1 illustrates a generalized block diagram of a computer system 20 in which a system according to the invention may be embodied. The computer system 20 includes a CPU 22 coupled to memory 24 and one or more peripheral devices 26 via a system bus 28. The system bus 28 carries data and control signals to the CPU 22, memory 24 and peripheral devices 26. The memory 24 preferably includes Random Access Memory (RAM), but may also be implemented with Read Only Memory (ROM), or a combination of RAM and ROM. The memory 24 stores data for one or more programs that may be executed in the computer system 20.

Figure 2 is a block diagram of a CPU 22 according to an embodiment of the invention. The CPU 22 includes multiple functional units 30, one or more register files 32, and an instruction unit 34. The register files 32 typically contain several general purpose registers 36 for storing values, addresses and possibly other data. The term "general purpose registers" can include floating point, fixed point, and branch target address registers, to name a few. In addition, the registers 32 in the register file 36 include a tag field 38. The use of the tag field 38 in speculative execution will be discussed in further detail below.

The architecture of the CPU 22 may vary. This particular architecture merely depicts the high level hardware design of a CPU 22 according to one possible embodiment. Speculative execution implemented according to the invention can provide performance improvement in a variety of CPU designs, including in particular, CPUs with multiple functional units or CPUs with multiple pipelined functional units. Speculative execution is particularly effective in enhancing performance in superscalar or Very Long Instruction Word (VLIW) computers.

In the process of running a program, the CPU 22 carries out a series of instructions stored in memory 24. The instruction unit 34 fetches an instruction from memory via the system bus 28 and then decodes the instruction. Depending on the type of CPU and/or the scheduling method used, an instruction may have more than one operation. The instruction unit 34 issues operations to a functional unit 30 or to multiple functional units (shown as stacked boxes in Fig. 2). The instruction unit 34 sends control signals to a functional unit 30 to carry out the operation or operations in an instruction. In response to these control signals, the functional unit 30 reads data such as address or value from the appropriate registers in the register file 32 and performs an operation. For some operations, the functional unit 30 writes a result back to the register file 32. For a memory store operation, the functional unit 30 reads a memory address and a value stored in the register file 32 and transfers the value directly to memory 24.

Figure 3 illustrates the format of an instruction 40 having multiple operations. The instruction includes three operations, each including an opcode and one or more operands. Starting from the left, the first operation in the instruction includes an opcode 42, a destination register field 44, and two source register fields 46, 48. The second operation has an opcode 50 and two source register fields 52, 54. Finally, the third operation has an opcode 56 and only a single source register field 58. The source register fields specify the location of the inputs to the operation, and the destination register field specifies the location of the result.

These three operations provide an example of a typical instruction. The first operation is an arithmetic ADD operation in which the functional unit 30 reads values stored in registers R1 and R2, adds them, and writes the result to reg-

ister R3. The second example operation is a memory operation 50 in which the first register R4 stores an address and the second register R5 stores a value. The functional unit 30 reads the address and value from the registers and stores the value in memory 24 at the specified address. The third example operation is a simple move (MOV) operation. To support speculative execution, the hardware semantics of this operation may be modified so that the operation may be used to check for deferred exceptions.

For example, the functional unit 30 reads the source register R6 and if an exception has occurred, initiates a branch operation to an error handling routine. The source register for the check operation is the same as the register to be checked for a deferred exception. The source register of the check operation may be the destination register of a speculative operation that has generated and deferred an exception, or of a speculative operation that has propagated the deferred exception. In either case, the source register that is checked contains information about the deferred exception.

A system for supporting speculative execution may be implemented in a computer system 20 as described with reference to Figs. 1-3. There are several alternative ways to implement a system according to the invention. While two specific embodiments will be described below, it should be understood that other variations are possible and will be apparent to those of ordinary skill in the art.

A first embodiment includes encoding speculative and non-speculative versions of operations in the instruction set architecture of the CPU. In this embodiment, the opcodes identify whether an operation is non-speculative or speculative. For example, the opcode 42 of the ADD operation in Fig. 3 could be encoded to designate the operation as either speculative or non-speculative. The version of the operations affects whether an exception generated by the operation will be reported or deferred. If a non-speculative operation generates an error, the functional unit 30 reports the error immediately. If a speculative instruction generates an exception, on the other hand, the functional unit 30 will defer reporting the exception. In the latter case, the exception will only be reported if the result of the operation is actually used in its home basic block.

A second embodiment provides a system for supporting speculative execution in a minimum instruction set architecture. In contrast to the first embodiment, this second embodiment supports speculative execution without including opcodes for speculative operations. Instead, the system treats a number of predetermined operations in an instruction set as speculative without requiring that operations be specially encoded. Like the first embodiment, this second embodiment also includes architectural support for deferring exceptions from speculative operations and for reporting these deferred exceptions.

In the second embodiment, the scheduler and CPU may treat a predetermined set of operations as speculative. This set of operations may potentially include any operation in the instruction set. However, operations whose effects are difficult to undo such as branch operations or operations visible outside the CPU should generally not be executed speculatively. All of the operations that are not visible outside the CPU may be treated as natively speculative operations by changing the hardware semantics in the functional unit 30. To support speculative execution in this embodiment, natively speculative operations are scheduled as speculative operations, and the CPU hardware includes support for deferring and reporting deferred exceptions from these operations. While a predetermined set of operations are treated as natively speculative in the CPU, operations visible outside the CPU remain non-speculative. An example of an operation visible outside the CPU is a memory store. It is considered to be visible outside the CPU because it interacts with memory outside the CPU, and more specifically, involves overwriting a location in memory 24. It is not preferable to execute such an operation speculatively because it may affect another process interacting with the same memory location. It is also more difficult to correct an error generated by an operation that affects the state of external hardware. For these reasons, it is preferable to treat operations visible outside the processor as non-speculative.

Using the approach of the second embodiment, an instruction set does not require speculative and non-speculative versions of operations. As a result, speculative execution can be supported in a manner that maintains compatibility with an existing instruction set architecture. The operations that are treated as natively speculative and non-speculative vary with the architecture of the CPU. In some systems, all operations except those visible outside the CPU can be treated speculatively. In one alternative, all operations could be treated as natively speculative. In another alternative, some operations could be natively speculative while others are provided as both speculative and non-speculative.

A system according to the second embodiment may be designed to be compatible with the binary form of a program, whether or not it has been optimized using speculative code motion. To take advantage of the speculative execution supported in the CPU, programs are optimized using speculative code motion to move operations for which the CPU has a speculative version. Two categories of programs may therefore exist: existing programs that have not been optimized using speculative code motion, and new programs that have been specifically optimized for the speculative execution supported by the CPU. To maintain compatibility with both programs, the CPU may include circuitry to support a non-speculative mode and a natively speculative mode.

If the CPU is to support both modes of operation, it must include circuitry for recognizing whether a program has been optimized for speculative execution. This circuitry may include a register such as the status word register where a mode bit may be stored. Before the CPU begins execution of a program, it reads this register and switches to the appropriate mode. The mode bit may be set in a variety of ways. For example, if a program has been compiled to take advantage

tage of speculative execution supported in the CPU, then the operating system may be programmed to set the mode bit. Alternatively, the compiler of the program may set the mode bit by identifying the mode in a status operation placed in the binary form of the program. Many other ways of specifying the mode of operation are possible and will be apparent to those of skill in the art.

While a system according to the second embodiment need not support two modes of operation, it is advantageous to support two modes of operation to maintain compatibility with existing programs. If the CPU is to execute only programs that have been specifically optimized for speculative execution, then the CPU need not support a non-speculative mode. An advantage of the second embodiment, however, is to maintain compatibility with existing instruction set architectures and programs that have already been compiled to run on CPUs based on these architectures. If compatibility with existing programs is important, then the CPU should support both a speculative and a non-speculative mode of operation.

The first and second embodiments include architectural support for deferring and reporting exceptions. This can include a tag field 38 in the registers of the register files. When an exception occurs in a speculative operation, the functional unit can set an error tag code in the tag field to indicate that an exception has occurred. This exception is not reported immediately, but rather, is deferred. Information about the exception may be written in the destination register to assist in error handling. This information can identify the instruction, and the operation within the instruction, which generated the exception. For example, the information may uniquely identify the program counter value of the operation that generated an exception. It may also identify the type of exception such as an address violation, arithmetic underflow and overflow, etc. If another speculative operation uses the result of this operation as an input, then the exception is propagated. To propagate the exception, an error tag code such as a tag bit in the tag field 38 is set in the destination register, and the exception information is copied into the destination register. In this manner, the system may defer exceptions generated by speculative operations until the result of the operation is actually used in the execution path of the program.

In one implementation, the data word includes a "data/status" field as well as a tag field 38. In the illustrated embodiment, the tag field 38 is outside of the conventional data field and indicates whether the data field is to be interpreted as correct data or as status. The status information describes either the nature of an error (error status) or it indicates how an operand resulting from a deferred exception may be correctly regenerated through exception processing (exception recovery status). In an alternative embodiment the tag field can be located in the data/status field. For example, IEEE floating point numbers include unused bits that can be used for this purpose.

If the operand is not tagged, the data is processed as normal data.

In order to support error reporting, the tag field (and possible error status) is propagated forward, by hardware, to a non-speculative operation which no longer propagates the error. If a speculative operation is executed and one or more input operands are tagged with error, its result operands are also tagged with error. When an operation is tagged as erroneous, it optionally carries an error status field indicating the nature of the error. The error status is propagated much like the tag.

In order to support deferred exception handling, the tag field (and possible exception recovery status) is propagated forward to a non-deferring use of the value which no longer propagates the deferred exception. Deferred exception propagation proceeds much like error propagation. If only one input is tagged with a recovery status its value is propagated as the operation result. If multiple operation inputs are tagged with a recovery status, any of the inputs may be selected as the result.

Error propagation takes precedence over propagation of deferred exceptions. Thus, if one input is tagged with error status, while another input is tagged with recovery status, error status is propagated.

Any information which is available at the time an exception is deferred and used at the time the exception is processed is encoded in the recovery status field. There are a number of types of information which may be put in the recovery status field depending on the exception recovery strategy. One example is the program counter of the instruction which originated the deferred exception. This supports recovery schemes which require knowledge of the instruction which caused an exception which was deferred. Some schemes, for example, use the recovery status to branch back to the location where the deferred exception originated.

Another example is an exception recovery status which encodes the operation needing re-execution to process a deferred exception. The status field can encode the original trapping instruction e.g. the instructions `rx=floating_add(ry,rz)` or `rx=memory_load(ry)` could be encoded in the recovery status.

An instruction can be re-executed non-speculatively, by decoding the recovery status, to process the deferred exception. This is done when a non-speculative operation processes the exception recovery status. Here, the exception status is decoded and the resulting instruction is executed non-speculatively. Processing resumes after processing the deferred exception. This scheme allows an exception to be deferred, but does not allow the result of the operation whose exception is deferred to again be used speculatively. Since the exception status includes data to support re-execution, this scheme does not require the use of compensation code.

The status field is especially useful in processing page faults. The status field can encode the load operation as noted above. Since the status field encodes the information necessary to process the page fault, there is no need to

generate fix-up or compensation code.

An error code in the tag field 38 is only one way of deferring an exception. Other hardware may also be used. For example, a speculative operation that generates an exception can write information to a buffer identifying that an exception has occurred and including exception information used in error handling. If the speculative instruction is actually used in the execution path of the program, then the error information can be read from the buffer, and the exception can be handled accordingly.

A system constructed according either the first or second embodiment of the invention also includes architectural support for reporting deferred exceptions. This support may include a check function incorporated into a non-speculative operation or may include the addition of a separate check operation. If operations visible outside the CPU are non-speculative, then a check operation could be incorporated into one of these operations. For example, as part of a memory store operation, the functional unit 30 could check for deferred exceptions. As described above, a memory store operation has two source registers - one that contains the datum to be stored in memory and one that contains the memory address. A memory store operation, when acting as a check operation, checks both of its source registers for deferred and/or propagated exceptions.

Referring to the example instruction in Fig. 3, the memory store operation could include exception detection semantics. As part of the memory store operation, the functional unit 30 would read the exception tag bit 38 of the source registers of the memory store operation to determine whether an exception has been deferred. If an exception has been deferred, the memory store operation would then execute a branch to an error handling routine.

As another example, a separate check operation could be created. This would involve encoding only one additional opcode for a check operation. These check operations may be inserted during the scheduling process. Check operations should be inserted so that each speculative operation that may generate or propagate an exception can be checked. The source register of the check operation may be the destination register of the excepting operation. Also, the source register may be a destination register of the last operation in a chain of propagated exceptions. To ensure that speculative operations are checked, check operations may be inserted after a speculative operation or after a chain of speculative operations. Because an exception is propagated to all operations that use the result of a speculative operation that has generated an error, the check operation need not be inserted for every speculative operation.

Referring again to Fig. 3, the third operation is an example of a check operation that may be used to check for deferred exceptions. This operation has only one source and no destination register. If an exception has been deferred, the source register of a check operation contains information that identifies the type of exception and operation that generated the exception. When an exception is detected, the information in the source register of the check operation may be used to handle the exception. Typically, when an exception is detected, the CPU branches to an exception handling routine.

It should be understood that the foregoing discussion describes one of many possible ways of detecting and reporting a deferred exception. As another alternative, a check operation could check an operation at a memory location to determine whether the operation at this location generated an error. For example, the check operation could check an operation in memory identified by its program counter address. The check operation may be encoded to check a program counter address such as "check operation at PC address 535" or "check the third last operation." This form of check operation could be an alternative to checking the destination register of a speculative operation. Since this check operation would not involve checking a destination register for a deferred exception, an additional buffer can be used to store information about the type of exception used to assist in error handling.

Figure 4 is a flow diagram of a method for supporting speculative execution according to an embodiment of the invention. This diagram illustrates the operation of a system for supporting speculative execution according to the first and second embodiments of the invention.

In the first step 62, the operations are scheduled. As set forth above, the operations may be scheduled either dynamically or statically, or perhaps using a combination of both. As part of the scheduling step 62, a check operation may be inserted into the home basic blocks of operations that have been moved above a conditional branch. In the first embodiment, the opcodes of speculative operations are encoded to distinguish them from a non-speculative version of the same operation. In the second embodiment, a number of predetermined operations are treated as speculative without being encoded as such. Once the program is scheduled, the CPU begins issuing operations for execution.

An operation is executed in the next step 64. In a CPU with multiple functional units 30 as shown in Fig. 2, operations may be issued to separate functional units simultaneously. If the functional units 30 are also pipelined, several operations may be at varying stages of execution at one time for each pipeline.

If an exception occurs during the execution of an operation (66), then the system reacts differently depending on the type of operation (68). For speculative operations, the exception is deferred (70). For non-speculative operations, however, the exception is reported and an error handling routine is initiated (72). Non-speculative operations are those operations that have not been moved above a conditional branch, but rather, remain in their home basic block. Depending on the implementation, non-speculative operations can include any type of operation. Examples of non-speculative operations include memory store operations that are not moved above a branch because they are visible outside the CPU, or a check operation that is inserted into the home basic block of speculative operations to check for a deferred



exception.

Depending on how the system is implemented, an operation is identified as speculative or non-speculative in different ways. In the first embodiment, speculative and non-speculative operations are differentiated by their opcodes. In the second embodiment, a predetermined number of operations are treated as speculative operations without requiring special encoding of the opcodes for these speculative operations. The CPU recognizes these natively speculative operations and defers exceptions generated by them. Similarly, the CPU recognizes the operations that are natively non-speculative and immediately reports any errors generated by them.

To maintain compatibility with existing programs, the CPU in the second embodiment may have two modes of operation: a speculative mode and a non-speculative mode. In the speculative mode, the predetermined set of operations are treated as natively speculative. Exceptions are deferred for these natively speculative operations. In the non-speculative mode, the predetermined set of operations are treated as non-speculative and exceptions are not deferred for these operations.

It is possible to treat a number of operations as natively speculative while also specifically encoding speculative and non-speculative operations. For example, all operations not visible outside the CPU could be natively speculative while all other operations such as a memory store could be provided in speculative and non-speculative versions. This approach is subject to the limitations of the particular instruction set architecture, which may or may not have enough flexibility to add additional opcodes.

The sequence of steps (64-70) include propagating a deferred exception. At the execution step 64, the functional unit 30 reads the error tag to determine whether an exception has occurred previously (66). If the operation is speculative (68), then the exception will be propagated by setting the error tag bit of the result register and transferring exception information to the result. Propagating an exception in this fashion, the system defers the exception (70) and reports it only if it is used in a non-speculative operation.

The sequence of steps (64-72) may include detecting and reporting a deferred exception. An exception is detected by executing a non-speculative operation (64) with error checking capability. Executing this type of operation includes checking for a deferred exception (66). The means for checking for a deferred exception vary depending on the implementation. Non-speculative operations may be implemented such that they check for a deferred exception by reading the error tag bit of an input or source register. For example, a memory store instruction may, as part of its semantics, read the error tag bit to check for a deferred exception and report an exception if one has been deferred. As another example, a check operation inserted into a home basic block may check for an exception by reading the error tag field of a result register or registers of speculative operations from that home basic block. If an exception has been deferred, the exception will be reported immediately (72). In this manner, the system for supporting speculative execution defers exceptions generated in speculative operations.

An increase in performance can be gained by deferring floating point exception processing. A floating point fix-up exception is invoked when the floating point hardware is unable to complete the computation of a correct result. For example, in most processors, the computation of a denormalized result of an IEEE floating point operation is too rare to justify additional hardware to perform the calculation. The hardware necessary for such a calculation is too complex and expensive to justify the cost.

In this case, it is preferable to compute the denormalized result in software to simplify the hardware. To perform this calculation in software, the program branches to a floating point fix-up exception code. The fix-up exception code is a program that computes the denormalized result.

Unfortunately, the processing of a denormalized result of an IEEE floating point operation in software decreases performance because it consumes several CPU cycles to complete.

A floating point fix-up exception should not be invoked during speculative execution because it negates the speed advantage obtained by speculatively executing operations. The problem is that during speculative execution, floating point fix-up exceptions may be invoked on behalf of an operation whose result is never used. This unnecessary processing is a costly waste of processor time. It is preferable, therefore, to branch to the fix-up exception code only if the result of the floating point operation is actually used in a non-speculative operation.

The following discussion provides one example describing how floating point exception processing can be deferred using a single register tag bit. Before beginning however, it is helpful to define some terms. An "a-net" is a re-executable network of operations. More specifically, an a-net is a multiple input single output network of floating point operations excluding any operation that cannot be executed speculatively (e.g. stores to memory). a-nets are restricted to acyclic graphs, meaning that no cycles are allowed.

All operations within the a-net are executed speculatively, and the processing of exceptions within the a-net is deferred until the final result of the a-net is checked using a non-speculative operation. The values of all inputs used in an a-net must be held in their corresponding input registers until such time as the output of the a-net has been checked for validity. Because these inputs are preserved, the operations in the a-net can be re-executed when it can be shown that the result is used by a non-speculative operation.

Processing of the exception is deferred using the tag field. If the tag bit within the tag field is set for an input operand in a speculative operation, then the error tag is also set in the result register of the speculative operation. As noted

above, the error tag bit determines whether the processor will interpret the data field as correct data or an exception. When an exception occurs the data field can include an exception code to assist in processing the exception. The exception code can include a status bit or bits to support re-execution of operations. One important use of a status bit is to indicate whether a fatal error has occurred. The exception code can also identify the nature of an error such as an IEEE underflow or overflow, or can identify the program counter value of the operation that generated an exception.

Either hardware or software can be used to propagate exception information. Hardware can be used to carry an error bit forward. In this case, the error bit is a special bit in the register dedicated to this function. The error tag bit is an example of such an error bit.

A software trap can be used to carry forward data in the data field of a register. This type of trap pushes the exception code through to the result register. In some situations, the trap for propagating this information may consume several processor cycles, but it is a better alternative than completely processing the floating point exception, which consumes considerably more cycles. If the error is fatal, there is no need to invoke the trap routine to propagate exception information. Thus, the status bit identifying a fatal error can identify operations where exception information should not be propagated with a trap.

Returning again to the example, a piece of fix-up code is compiled for each a-net. The fix-up code non-speculatively computes the output for the a-net as a function of all inputs to the network. This is performed by copying all of the operations within the a-net into the fix-up code and using non-speculative versions of these operations. The fix-up code is terminated using a return from sub-routine instruction. It should be noted that the sequence of operations in the fix-up code can include speculative and non-speculative operations. However, the sequence of operations in the fix-up code are executed in a non-speculative manner: non-speculative operations are executed and speculative operations are not. For example, speculative operations in the fix-up code may include a flag or a reference to a flag that causes the CPU to avoid executing the speculative operations in the process of executing the fix-up code. In one specific implementation, the opcode of a speculative operation in the fix-up code may include a reference to a flag called a "predicate", which indicates whether the operation should be executed. When the fix-up code is executed in a non-speculative manner, the CPU checks the predicate referred to in the opcode and determines whether or not to execute the operation. To avoid executing a speculative operation, the reference in the opcode points to a predicate flag that instructs the CPU to avoid executing the operation. There are alternative ways to implement this type of a predicate. In addition, it is also possible that the fix-up code will contain only non-speculative operations.

The result of the a-net may be used along multiple program paths. On each program path, the first non-speculative use of the output result from an a-net may require the execution of unique compensation code corresponding to the given a-net. Each such non-speculative use is set in accordance with the address indicating the beginning location of the unique fix-up code for that a-net. The correspondence between the beginning location of the unique fix-up code for the a-net and the non-speculative use is called the "fix-up map." A fix-up map may be maintained in software using techniques such as a hash table.

When the program non-speculatively uses the result operand of an a-net, an exception is processed if the operand error tag bit is set. If the error tag bit causes the non-speculative operation to trap, the instruction does not execute. Instead, the hardware jumps to an exception handler which processes exceptions. The exception handler sets a return address so that the fix-up code will resume execution at the beginning of the trapping operation. The exception handler identifies the trapping operation and the corresponding tagged operand and uses these as a key to consult the fix-up map. The fix-up map provides an address identifying the origin of a fix-up code sequence.

The exception handler then jumps to the fix-up code sequence. Since the fix-up code is executed non-speculatively, it either causes a deferred exception to be successfully processed, or it results in a fatal error. After the fix-up code is executed and the result is correctly calculated, execution resumes at the instruction which caused the fix-up trap. After execution resumes, the a-net has been "checked". In other words, all operands used and generated within the network are known to be correct.

In general, it is possible for one a-net to have inputs which correspond to either intermediate results or outputs from a previous a-net. This type of network is termed a "dependent" network. On each path through the program, the result of a dependent network must be checked after the network upon which it depends has already been checked. This rule is enforced to guarantee that all inputs to an a-net are checked and are correct prior to re-execution.

Table 1 below provides an example program to illustrate the process of deferring the calculation of denormalized results of an IEEE floating point operations. In this example program, all operations are floating point operations.

TABLE 1

/\*previous code\*/

if (rx) branch to w

r5=op1(r1,r2)

r6=op2(r3,r4)

r7=op3(r5,r6)

r8=op4(r7,r8)

r9=op5(r7,r8)

if (r9) branch to x

r10=op6(r8,r12)

r11=op7(r7,r10)

if (r11) branch to y

In some cases, it may be important to speculate most of the operations without incurring penalties due to unnecessary exceptions. This example assumes that branches are non-speculative consumers of program results (branch conditions). The use of an operand within a branch causes that operand to be checked. If the tag bit for the operand is set, compensation code for the a-net which computes the result is executed.

The example program in Table 1 is in the form of a superblock. A superblock is a linear code sequence with a single entry at the top and possibly multiple exits, each marked by a branch within the sequence.

The code in Table 2 below shows how the superblock of Table 1 is modified by performing speculative code motion. Operations with an asterisk are speculative operations which have been moved above a branch within the superblock.

TABLE 2

| Operation | /*previous code*/    |
|-----------|----------------------|
| 1         | r5=op1*(r1,r2)       |
| 2         | r6=op2*(r3,r4)       |
| 3         | r7=op3*(r5,r6)       |
| 4         | r8=op4*(r7,r8)       |
| 5         | r9=op5*(r7,r8)       |
| 6         | r10=op6*(r8,r12)     |
| 7         | r11=op7*(r7,r10)     |
| 8         | if (rx) branch to w  |
| 9         | if (r9) branch to x  |
| 10        | if (r11) branch to y |

Fig. 5 shows two a-nets, a-net one (80) and a-net two (82), for which fix-up code is generated. a-net one is checked when operation 9 checks its branch condition operand (r9) (84). a-net two is checked when operation 10 checks its branch condition operand (r11) (86).

The fix-up map is used to establish correspondence between the occurrence of an exception and the location of the fix-up code which reprocesses the exception. The fix-up map for this example is provided in Table 3 below. The fix-up map is a table including a column for an excepting operation, the operand position, and the compensation location. The operand position identifies the operand in the operation that uses the result of a speculative operation. The compensation location identifies the location of a compensation code sequence.

TABLE 3

| excepting instruction | operand position | compensation location |
|-----------------------|------------------|-----------------------|
| 9                     | operand 1        | 100                   |
| 10                    | operand 1        | 200                   |

Table 4 below illustrates an example of the compensation code for this example. The compensation code for each of the a-nets is reached through the fix-up map which indicates the beginning location of a compensation code sequence.

TABLE 4

|     |   |
|-----|---|
| 100 | r5=op1(r1,r2)<br>r6=op2(r3,r4)<br>r7=op3(r5,r6)<br>r8=op4(r5,r6)<br>r9=op5(r7,r8)<br>return |
| 200 | r10=op6(r8,r12)<br>r11=op7(r7,r10)<br>return  |

It should be apparent to those of skill in the art that this exception processing can be extended to other types of exceptions such as deferred processing of a page fault. Just as calculating a denormalized result for a floating point operation is costly in terms of processor cycles, the processing of a page fault is also time consuming because it requires loading data from secondary storage. Using the technique outlined above, the processing of a page fault can be deferred until it is certain that the data from a load operation that generated the page fault is used non-speculatively.

Though the system and methods of the invention have been described in detail in the context of alternative embodiments, it should be understood that the invention may be implemented in a variety of ways without departing from the scope of the invention. For example, the instruction set architecture can include non-speculative and speculative versions of operations, natively speculative operations, or a combination of both. The means for deferring an exception may vary depending on the CPU architecture. For example, a single error tag bit may be added to general purpose registers in a register file, or perhaps a buffer could be used to keep track of deferred exceptions. The means for checking for deferred exceptions may vary as well. As described, the hardware semantics of a CPU can be designed to check for deferred exceptions when one or more types of non-speculative operations are executed. Alternatively, a separate check operation could be added to the instruction set architecture.

The support for re-execution of operations can vary as well. For example, unused bits in the data field of a register can be used to store a status code. For instance, the IEEE specification for floating point numbers leaves several bits unspecified in some cases. These unused bits can be used to store a status code. Many other variations to the embodiments are possible without departing from the scope of the invention.

In view of the many possible embodiments to which the principles of our invention may be put, it is emphasized that the detailed embodiments described herein are illustrative only and should not be taken as limiting the scope of our invention. Rather, we claim as our invention all such embodiments as may come within the scope and spirit of the following claims and equivalents thereto.

## Claims

1. A method for processing speculative operations in a processor comprising:

issuing a speculative operation in a processor;  
 if execution of the speculative operation generates an exception, then setting a tag code in a tag field of a result register of the speculative operation to defer the processing of the exception;  
 while executing a non-speculative operation, checking whether the tag code is set in an input register of the non-speculative operation;  
 and if the tag code is set, then processing the exception and re-executing a non-speculative version of the speculative operation.

2. The method of claim 1 further comprising:

if the speculative operation generates an exception, then setting an exception code in the result register of the speculative operation; and  
 propagating the exception code if the result register of the speculative operation is an input to another speculative operation.

3. The method of claim 2 wherein the propagating step includes propagating the exception code with a trap routine.

4. The method of claim 1 further comprising:

if the speculative operation generates a fatal error, then setting a status bit in the result register indicating that a fatal error has occurred.

5. The method of claim 1 further comprising:

if the speculative operation generates the exception, then encoding an operation requiring re-execution in status bits in the data field of the result register; and  
 re-executing the operation requiring re-execution by decoding the status bits and re-executing the operation requiring re-execution non-speculatively.

6. The method of claim 1 comprising the step:

during compilation of a program, identifying a sequence of operations to be executed speculatively and maintaining an alternative version of the sequence of operations that includes non-speculative operations; and  
 if one or more of the sequence of speculative operations generates an exception, then re-executing the alternative version of the sequence of operations in a non-speculative manner.

7. The method of claim 1 wherein the non-speculative operation executed during the checking step is a check operation, and executing the check operation includes reading the tag bit in the result register of the speculative operation to detect whether an exception has been deferred.

8. The method of claim 1 wherein the speculative operation is a floating point operation and processing the exception includes calculating a denormalized result of the floating point operation.

9. The method of claim 1 wherein the speculative operation is a load operation that generates a page fault and processing the exception includes executing a non-speculative load operation.

10. The method of claim 1 wherein the step of re-executing includes re-executing a sequence of non-speculative operations.

11. A method for processing speculative operations in a processor comprising:

issuing a speculative floating point operation in a processor;  
 if execution of the speculative floating point operation requires calculating a denormalized result, then setting a tag bit in a result register of the speculative floating point operation to defer calculation of the denormalized result;  
 while executing a non-speculative operation, checking whether the tag bit is set in an input register of the non-speculative operation;  
 and if the tag bit is set, then calculating the denormalized result of the speculative floating point operation including re-executing a non-speculative version of the speculative floating point operation.

12. A system for processing speculative operations comprising:

a register file including general purpose registers, the general purpose registers including a tag bit;  
a functional unit in communication with the register file for executing operations, the functional unit in communication with the register file for setting the tag bit when a speculative operation generates an exception, for reading the tag bit to detect a deferred exception, and for re-executing a non-speculative version of the speculative operation if the deferred exception is detected.

5

10

15

20

25

30

35

40

45

50

55

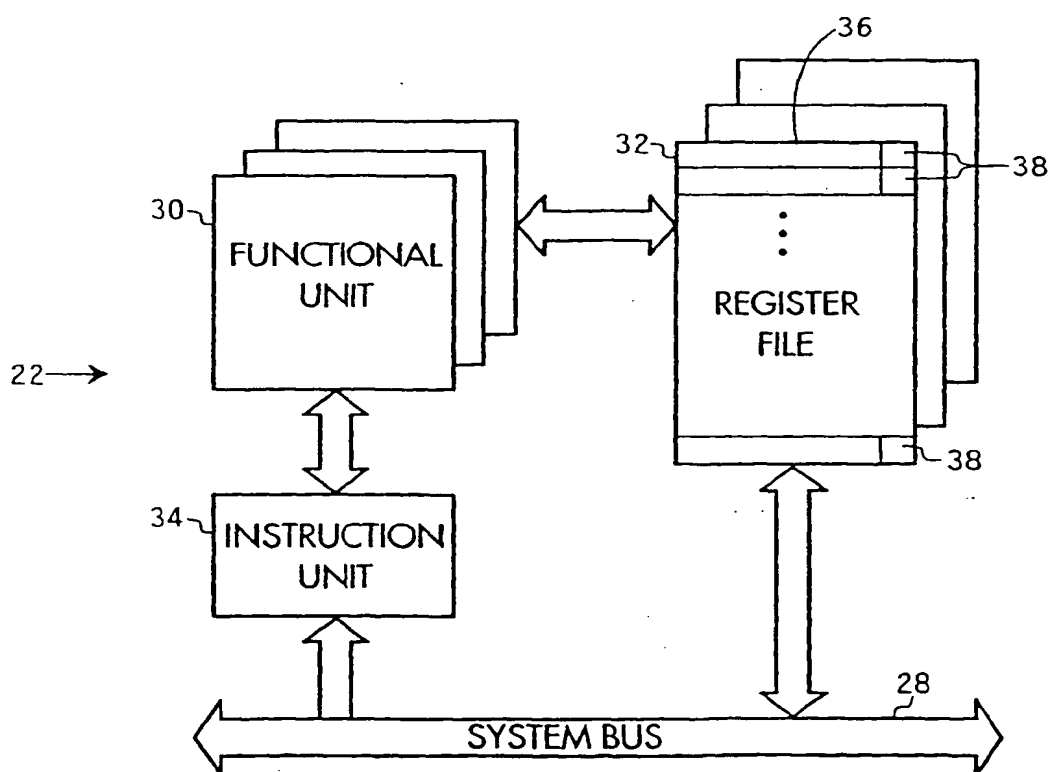
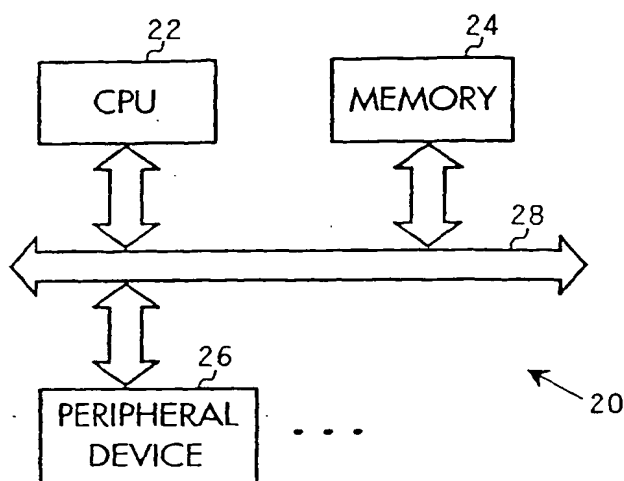


FIG. 3

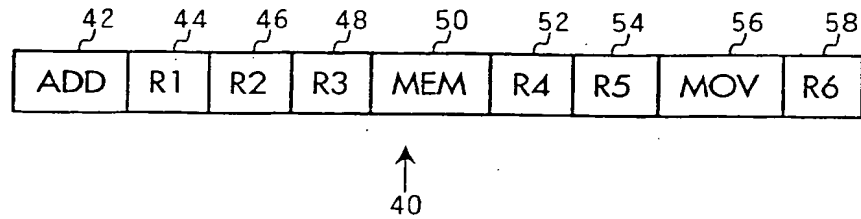


FIG. 4

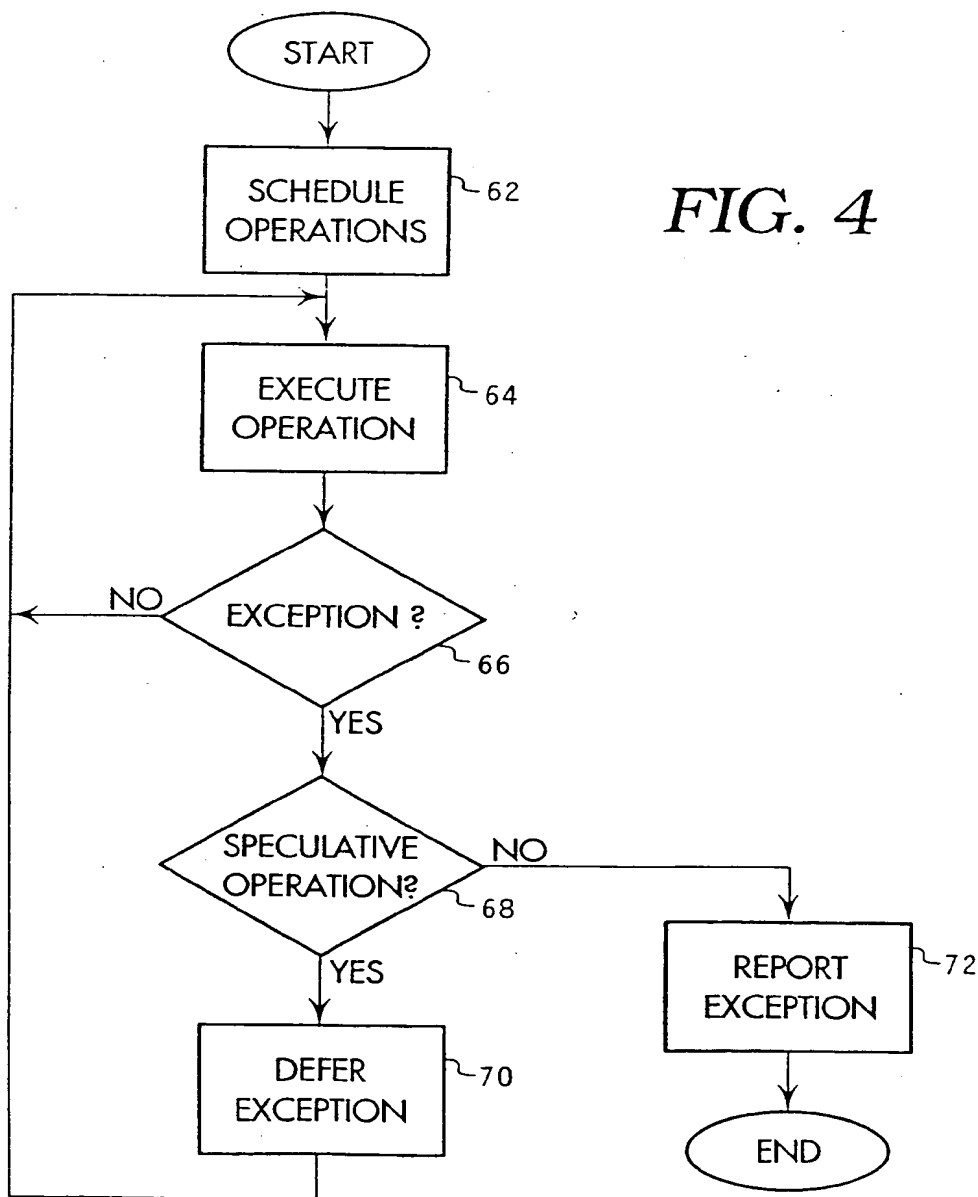
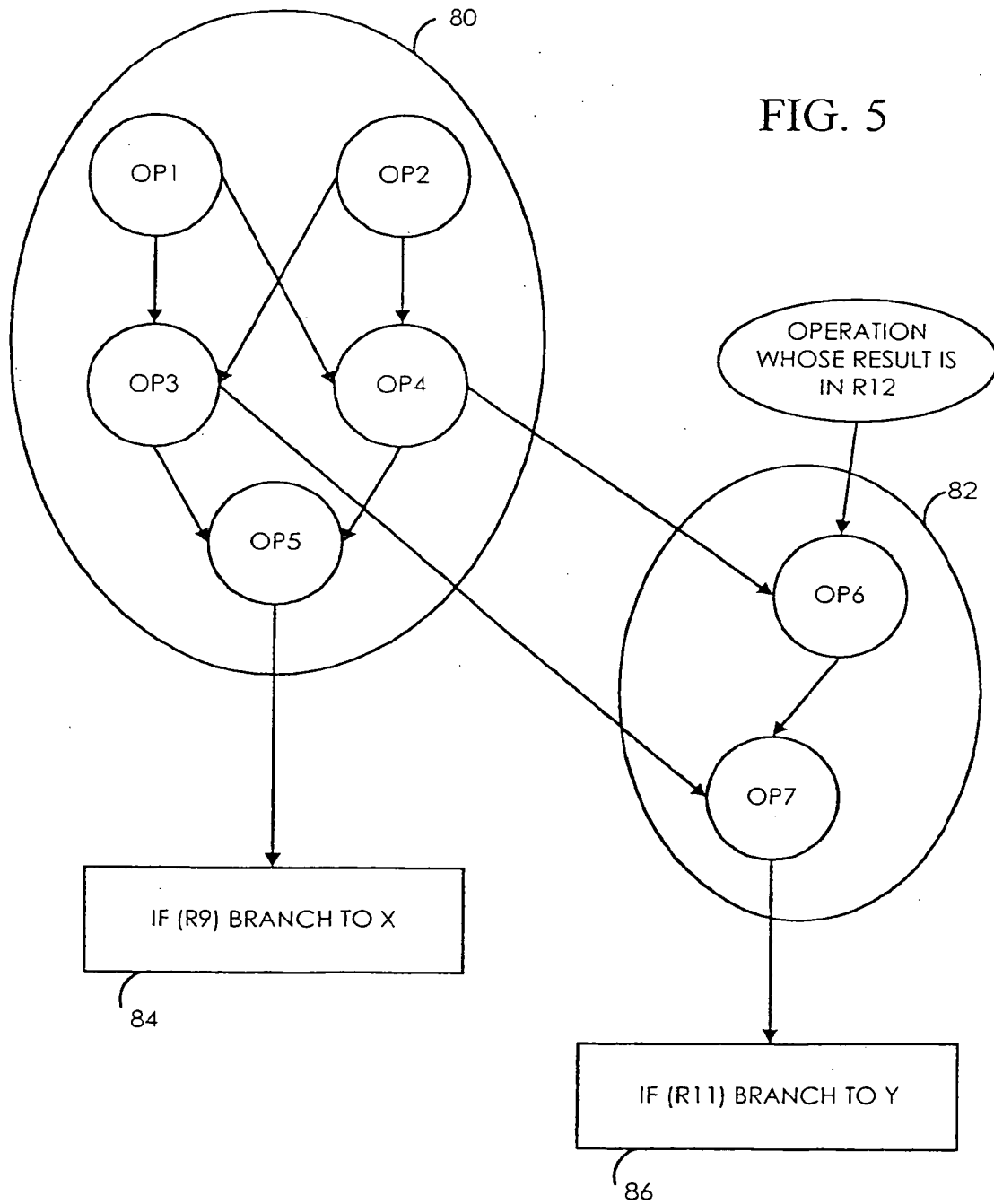




FIG. 5





European Patent  
Office

# EUROPEAN SEARCH REPORT

Application Number

EP 97 10 1810

| DOCUMENTS CONSIDERED TO BE RELEVANT  |   |                   |  |
|--|---|-------------------|--|
| Category   | Citation of document with indication, where appropriate, of relevant passages         | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
| X  | US 5 421 022 A (F. X. MCKEEN ET AL)<br>* the whole document *                         | 1-3,7-12          | G06F9/38                                     |
| A  | ---   | 4-6               |  |
| D,P,<br>X  | GB 2 294 341 A (HEWLETT-PACKARD COMPANY)<br>* the whole document *                    | 1-3,7-9,<br>11,12 |  |
| A  | US 5 428 807 A (F. X. MCKEEN ET AL)<br>* the whole document *                         | 2,3               |  |
| P,A  | WO 96 23254 A (INTERNATIONAL BUSINESS MACHINES CORPORATION)<br>* the whole document * | 1-12              |  |
| The present search report has been drawn up for all claims   |   |                   | TECHNICAL FIELDS SEARCHED (Int.Cl.6)         |
|  |   |                   | G06F   |
| Place of search  | Date of completion of the search  | Examiner          |  |
| BERLIN   | 10 April 1997   | Abram, R          |  |
| <p><b>CATEGORY OF CITED DOCUMENTS</b></p> <p>X : particularly relevant if taken alone<br/> Y : particularly relevant if combined with another document of the same category<br/> A : technological background<br/> O : non-written disclosure<br/> P : intermediate document</p> <p>T : theory or principle underlying the invention<br/> E : earlier patent document, but published on, or after the filing date<br/> D : document cited in the application<br/> L : document cited for other reasons<br/> &amp; : member of the same patent family, corresponding document</p> |   |                   |  |

EPO FORM 1503 (03.92) (POMC01)